CHAPTER 4

Value-based purchasing: Pay for performance in home health care

CHAPTER 4

Value-based purchasing: Pay for performance in home health care

Chapter summary

In the Deficit Reduction Act of 2005, the Congress asked the Commission to discuss the design of a pay-for-performance (P4P) system in home health care as part of a broad set of initiatives to improve the value of health care that Medicare purchases. Providing financial incentives for quality is one tool the Medicare program can use in home health and other settings. P4P should be used in tandem with other payment reforms (e.g., increasing the accuracy of payments) as well as with other quality incentives (e.g., public reporting).

The first key decision in the design of a P4P system is how to fund the reward pool. As a principle, the Commission has stated that P4P should be budget neutral, neither adding nor removing money from the system. Thus, the system could be funded by redistributing payments from poor performers to high-quality performers and to providers who are improving.

Another set of key decisions involves how to set thresholds for performance. One way to set a threshold is to predetermine a

In this chapter

- Pay for performance in Medicare: The Commission's design principles
- Pay for performance for home health
- Circumstances of the home health sector
- Additional technical information on home health pay for performance

percentage of providers (e.g., rewarding the top 10 percent of providers). Another alternative is to choose a minimum score and use a test of statistical significance: High performance is a score statistically significantly above the average, poor performance is a score statistically significantly below the average, and improvement is a score statistically significantly greater than the provider's previous score.

A system that rewards both attainment of high quality and improvement toward high quality must find a balance between the two rewards. If the rewards are exclusive (a provider can receive either an attainment reward or an improvement reward but not both) then less weight could be placed on the improvement rewards since those providers are, by definition, providing lower quality care as measured by the P4P system.

A final decision in P4P design is to determine the size of the reward. In a budget-neutral system, the size of the reward is constrained by the size of the penalty placed on poorly performing providers. One implication of the Commission's principle that P4P should be budget neutral is that when money is removed from the system to fund the pool, then the entire reward pool should be spent on rewards. The size of the reward for the provider should be a percentage of the provider's Medicare payments.

The circumstances of home health care may pose some challenges for P4P in that sector. The payment system has some inaccuracies, and payments have been more than adequate. The Commission will continue to consider reforms to the payment system. P4P should be put in place at the same time Medicare improves the payment system to create stronger incentives to improve quality. ■

MedPAC recommended that Medicare build financial incentives for quality into payments to hospitals, physicians, home health agencies (HHAs), dialysis providers, and Medicare Advantage plans (MedPAC 2004, 2003). Medicare's current payment systems are neutral or negative toward the quality of services; these systems do not promote the program's goals to provide high-quality services to its beneficiaries and to be a good steward of public resources. The program should link payment to quality through a pay-for-performance (P4P) program to increase the value of health care spending. P4P should be used as one payment policy tool along with reforms that address other weaknesses in the payment system and other incentives for quality.

The Congress asked the Commission to address several key design issues in developing a system that links payment to performance in home health care as part of a broad initiative to encourage value-based purchasing in the Medicare program. The Deficit Reduction Act of 2005 requested this mandated report (see text box, p. 80). The mandate posed four questions: How should P4P be funded? What is the threshold for a reward? How should improvement and attainment be balanced? What is an effective size for the reward?

Pay for performance in Medicare: The Commission's design principles

The Commission has developed principles to guide the design of a P4P program and to select the quality measures that would support it.

Program design features

The Commission calls for P4P programs that:

- Reward providers based on attaining or exceeding certain benchmarks *and* improving at certain benchmarks. This principle seeks to encourage as many providers as possible to improve, thus maximizing the benefit of the program to as many beneficiaries as possible. Providers already performing at high levels will be rewarded for their efforts. Those who score low at baseline will have an incentive to improve. If all providers improve over time, improvement incentives can be phased out of the system.
- Are funded by setting aside a small proportion of the current payment—initially 1 percent to 2 percent. The first dimension of this principle is whether the policy should be funded by withholding dollars or whether new spending is necessary.[1] Through a separate process, the Commission evaluates the adequacy of payment levels every year when it recommends payment updates for providers. The Commission determined that the P4P initiative should be funded within current levels of spending. The primary rationale was to shift the incentives of payment, not the level.

  The second dimension is whether the size of the incentive is enough to encourage provider change or whether it is too disruptive. Evidence about the "right" level for incentives is limited.[2] In a budget-neutral program, smaller incentives may be more powerful as providers perceive the penalty dollars as lost income. The much smaller 0.4 percent incentive for hospitals called for by the Medicare Prescription Drug, Improvement, and Modernization Act of 2003 was designed to encourage data reporting as a condition for receiving a full update; there was a penalty for nonparticipation. It resulted in nearly universal hospital reporting on certain process measures.

  Others have suggested that, if the dollars are withheld, even 1 percent to 2 percent could be significant and potentially harm providers that may be at low levels of quality. This concern was one rationale for suggesting that improvement from low levels should also be rewarded.

  Given the limited evidence on the right level, and to ensure minimal disruption for beneficiaries and providers, the Commission chose to recommend that 1 percent to 2 percent be set aside, at least initially. The Commission expects the percentage to increase as the Medicare program and providers gain more experience with P4P.
- Distribute all payments that are set aside to providers that meet reward criteria.
- Establish a process for evolution of the program, together with private purchasers and other public purchasers. The P4P design should be evaluated and changed over time. This system should be a learning system.

Mandate for report

The Deficit Reduction Act of 2005

MedPAC Report on value based purchasing.

Not later than June 1, 2007, the Medicare Payment Advisory Commission shall submit to Congress a report that includes recommendations on a detailed structure of value based payment adjustments for home health services under the Medicare program under title XVIII of the Social Security Act. Such report shall include recommendations concerning the determination of thresholds, the size of such payments, sources of funds, and the relationship of payments for improvement and attainment of quality. ■

Criteria for quality measures for a pay-for-performance program

Based, in part, on the experiences of private-sector initiatives, the Commission developed criteria for determining whether the measures and measurement activities for each provider setting were sufficient to distinguish between high- and low-quality performance. These criteria are:

- Well-accepted, evidence-based measures must be available. They should be accepted by independent quality experts and should be familiar to providers. While few individual measures are perfectly valid or reliable, they should identify real differences in provider quality.
- Collecting and analyzing data should not be unduly burdensome for either the provider or CMS. To minimize the burden of collection and analysis, CMS should base quality measures on data it currently collects, wherever possible. The need for additional information should be balanced against the value of the information to the provider being measured, patients, and the Medicare program.
- Incentives should not discourage providers from taking riskier or more complex patients. Appropriate risk adjustment is always important when comparing provider quality. To address this concern, the program could use measures that—in general—are not affected by the complexity of the patient, such as process, structure, and patient-reported experience of care measures. Risk adjustment is critical for outcomes-of-care measures.
- Most providers should be able to improve on the available measures. This criterion has several dimensions. For one, the measures should capture aspects of care the providers can affect. Another dimension is that the measures should be related to aspects of quality that most need improvement; there should be room for real gains in quality. Another dimension is scope. The measures should apply to a broad range of care and providers; the greater the proportion of providers whose care is measured, the broader the impact will be on beneficiaries. It is also important to measure a broad range of the types of care delivered in the setting. Measures focused on specific conditions are already available in most settings, but to capture a broad range of care in each setting, measures that apply to all types of patients (e.g., safe practices, use of patient registries, and patient perceptions of care) should be added over time. A starter set of measures could satisfy this criterion and not necessarily encompass all care, all providers, and all patients.
- A P4P measure set should evolve to become more comprehensive. Ideally, measures should also reach across settings to align incentives across providers such as hospitals, skilled nursing facilities, and physicians working together to reduce readmissions to acute care hospitals. After Medicare chooses an initial measure set, CMS will need to alter, add, and drop measures and ensure that research is under way to create or validate other measures. A single entity could help coordinate public and private efforts and, based on the advice of quality experts, make recommendations on measures.

What will make pay for performance work?

Providing incentives for quality can increase value by prompting providers to begin addressing the current shortcomings of health care.[3] Results such as the high level of evidence-based care for cancer in the first year of the United Kingdom's physician pay-for-performance (P4P) program (Doran et al. 2006), the increase in cholesterol screening during California's physician P4P program (Integrated Healthcare Association 2006), and patients receiving aspirin after a heart attack under CMS's hospital P4P demonstration provide evidence that providers respond to incentives to improve their performance, increasing the quality of health care.

The Agency for Healthcare Research and Quality (AHRQ) synthesized economic, psychological, decision, and organizational theories to describe other factors that could lead providers to respond to—or ignore—a P4P program (Dudley et al. 2004). We summarize these factors in this text box.

Providers are more likely to respond to financial incentives if expected revenue is greater than or equal to costs. If the direct costs and opportunity costs of responding to the incentive outweigh the financial return, then the incentive is likely to fail. However, this may be mitigated by some of the nonfinancial incentives, such as a commitment to professionalism, the mission of the organization, and the provider's potential loss of standing among peers or in the community (Town et al. 2004). These "costs," in terms of the provider's reputation, will be greater if the P4P information is widely available.

Providers who think they have greater control over what is measured will have a greater response. For this reason, structural and process measures may generate a greater response than outcome measures.

Providers under fee-for-service payment are more likely to respond to incentives to produce more units of service—more discharges or more episodes of home health care—because improving quality in a way that increases use of services increases revenue. Alternatively, providers in a capitated payment system may be less attracted to incentives that require more services to be provided within the bundle of payment.

Researchers at the University of Minnesota expanded on AHRQ's list with provider characteristics that will affect a provider's response to P4P (Town et al. 2004). For example, providers that are risk averse will respond more strongly to avoid a penalty.

If different payers coordinate their efforts, P4P is more likely to succeed because providers can receive consistent incentives and avoid duplicative or incompatible requests for quality data. Also, the coordination of effort leads to a greater impact by capturing a larger portion of providers' total revenue. ■

Pay for performance for home health

In this section, we apply the Commission's general principles to the specific challenge of developing a Medicare P4P system for HHAs. We use an illustration of a home health P4P system to discuss the decisions to be made at each point. This illustration is only one of many possible designs for a P4P system; factors that influence whether P4P is likely to have an impact on quality should also be considered (see text box). Our use of a single model is for the benefit of clarity and does not imply an endorsement of this particular set of design choices. We chose six real agencies; using their actual quality and financial information from 2005, we present the rewards and penalties that would accrue in a system that pays more for high-quality care and less for low quality.

There are several decision points in the design of home health's P4P system. At each of these points in the model, we discuss the alternatives to the path we chose for the purpose of this illustration. The major decision points are:

- funding the reward pool
- measuring agency quality
- setting thresholds for reward and penalty
- balancing improvement and attainment
- calculating the rewards

For the purposes of illustration, we discuss a model that funds the reward pool by withholding 5 percent of payments from each HHA. While this is not the only design consistent with the Commission's principles, it is provided to illustrate one possible configuration of P4P in home health care. The model uses a quality measure based on improving or stabilizing functional outcomes and avoiding potentially preventable unplanned hospitalizations and trips to the emergency room (ER). To determine whether an agency will be rewarded or penalized, its quality score is compared to a national benchmark level of quality (the threshold) to determine whether it is statistically significantly higher or lower than the benchmark. The model also includes a measure of the agency's improvement in quality. The reward for attaining high quality is twice as large as the reward for improvement in this model. Rewards and penalties are calculated as a percentage of the agency's Medicare payments. We also discuss additional design features, such as addressing agencies with few patients and ways to improve the P4P system and the quality measures over time.

Funding the reward pool

The first decision is how to fund the reward pool. This involves two issues: (1) whether the funding should be budget neutral, new money, or from savings elsewhere in the program; and (2) how much funding should go to payment for performance.

Source of funding

The Commission has stated as a principle that P4P should be budget neutral. In a report on rewarding provider performance, the Institute of Medicine (IOM) also recommended a budget-neutral funding source (IOM 2006). The model applies budget neutrality by withholding 5 percent of Medicare revenue from poor performers to fund the reward pool for high performers. Thus, the reward and the penalty pools redistribute spending within the home health sector and do not add new money to it.

A P4P system that includes potential penalties (which is implicit in a budget-neutral program) may be more powerful than a system with the same percentage of payment without penalties because economic actors assign more value to potential income lost than to rewards won (Kahneman and Tversky 1979). If providers are at risk for losing revenue, then low-quality providers could perceive even 1 percent to 2 percent of payments as significant.

In contrast to the Commission's design principle, CMS uses savings generated by home health quality improvement in other sectors of Medicare to fund rewards for HHAs in its proposal for a demonstration. The demonstration would increase the amount of spending in the home health sector but would not increase Medicare spending as a whole because spending would be reduced in other sectors. Under the demonstration, if the HHAs in the demonstration keep their patients out of the hospital more often than agencies outside of the demonstration, then the amount saved on hospitalizations avoided will be available as rewards to high-quality HHAs that participate in the demonstration. If savings are not achieved, then no money will be available for rewards.

If a program were funded based on savings, IOM observed that it would not be possible to predict the size of the reward pool until the experience for the entire year in multiple sectors is gathered and analyzed, creating a long lag between implementing the program and rewarding providers and resulting in instability from year to year. IOM also noted that it would be difficult in a generated savings funding system to attribute spending decreases in one sector (e.g., hospitals) to quality interventions in a different sector (e.g., HHAs). This challenge would be compounded if and when P4P systems in different sectors are running simultaneously. For example, if both home health and skilled nursing facility P4Ps were running, the program should not "spend" the hospital savings twice, even though improvements in both skilled nursing facilities and home health care might have contributed to reduced hospitalizations. This funding source is likely to be unstable because it might be difficult to generate increasing savings year after year.

Providers may not perceive a funding system based on savings to be fair if improvements in their quality do not generate savings in other sectors. Providers may also perceive the complicated calculation of savings to be inaccurate. Finally, there may be a "free rider" problem if the savings some exemplary providers generate are attributed to all.

TABLE 4–1 **The pay-for-performance model withholds 5 percent of Medicare payments**

	Agency					
	1	2	3	4	5	6
Total Medicare payments	$192,000	$755,000	$4,706,000	$2,106,000	$415,000	$764,000
Payment withheld	$9,600	$37,700	$235,300	$105,300	$20,800	$38,200

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

A positive attribute of funding based on savings is its explicit link between high quality and resource use in achieving greater efficiency. It may appeal to policymakers because it builds an explicit incentive to generate savings for Medicare into the P4P program. If such a system were effective, one might imagine a future phase of the program in which Medicare keeps some of the savings and thus lowers total Medicare spending. Finally, such a system allows the program to fund a reward pool without penalizing (and presumably antagonizing) providers who participate in Medicare voluntarily or seeking new money from outside the program.

Level of funding

The Commission recommended starting P4P with a small portion of payment. Evidence on the right level for incentives is limited (Rosenthal et al. 2005). One survey of private-sector efforts found that purchasers report needing incentives of 5 percent to 20 percent to influence the behavior of physicians and 1 percent to 4 percent to influence hospitals. Applying these findings to a program as large as Medicare is problematic. We do not know what portion of providers' overall payment these percentages represent. Because Medicare payment often represents a higher percentage of a provider's total revenue than does a single private payer, a smaller percentage of Medicare's payment may be enough to encourage change. In CMS's Premier hospital demonstration, preliminary results show improvement in all conditions in the first four quarters in anticipation of financial rewards of either 1 percent or 2 percent for those in the upper rankings (Premier 2006).[4] The Commission expects the percentage to increase as the Medicare program and providers gain more experience with P4P.

As a general guide, the Commission suggested that P4P programs begin with 1 percent or 2 percent of payments. The model withholds 5 percent of payments. One could view the model as a program that started with a smaller withhold and grew over several years to the 5 percent level. In 2005, Medicare payments for home health services totaled $12.5 billion. The 5 percent withhold would generate $625 million in the pool for rewards. Annual Medicare payments to individual agencies ranged from about $125,000 to $6.5 million.[5] At the agency level, a 5 percent withhold would amount to a payment reduction ranging from $6,300 for some of the smallest agencies to $325,000 for some of the largest. The median agency received $1 million in Medicare payments and would have a withhold of $50,000.

Illustration of a home health P4P model

For illustrative purposes, the model (Table 4-1) withholds 5 percent of revenues from six agencies to demonstrate the reward pool.

Measuring agency quality

The core of home health quality measurement is the 31-measure Outcome-Based Quality Improvement (OBQI) set. CMS developed the OBQIs to use in their public reporting of HHA quality and to track changes in quality over time. The OBQI set includes the measures of outcome, stabilization, and improvement shown in Table 4-2 (p. 84).

These measures are based on comparison of patients' level of function at the beginning and end of their home health treatment as measured by the Outcome and Assessment Information Set (OASIS) patient assessment tool. Most patients can be included in most measures.

TABLE 4–2 **OBQI measure set**

Outcome	Stabilization	Improvement
• Acute care hospitalization • Any emergency care provided • Discharge to community	Stabilization in: • Bathing • Grooming • Transferring • Light meal preparation • Laundry • Housekeeping • Shopping • Telephone use	Improvement in: • Bathing • Grooming • Transferring • Light meal preparation • Laundry • Housekeeping • Shopping • Telephone use • Ability to dress lower body • Ability to dress upper body • Ambulation • Bowel incontinence • Confusion frequency • Dyspnea (shortness of breath) • Eating • Frequency of pain • Management of oral medications • Toileting • Urinary incontinence • Urinary tract infection

Note: OBQI (Outcome-Based Quality Improvement).

CMS has used about a dozen of these measures to assess individual HHAs' quality for the past several years on the Home Health Compare website. These measures satisfy most of the Commission's criteria for use in P4P: They are valid, reliable, generally accepted by researchers, and familiar to providers.[6] Providers can improve on these measures. They are derived from data that are routinely collected from HHAs and processed by CMS; they do not pose a new data burden.

A composite quality score

For illustrative purposes, we used a quality score that combines 20 home health outcomes into a score called the Standardized Quality Index (SQI). Additional technical information is provided at the end of this chapter. The SQI includes patients who improve at activities of daily living as well as those whose level of functioning is stable. It includes penalties for potentially avoidable hospitalizations and potentially avoidable use of the ER, both of which indicate lower quality and suboptimal resource use. The SQI groups patients into categories by their primary diagnosis. The measurement is restricted to patients for whom Medicare is the primary payer.

The SQI gives agencies credit for stabilizing patients who do not improve. This allows the system to capture the quality of care provided to patients who use home health care to remain safely at home, stabilize their condition, and avoid institutional care settings such as a nursing home.

The score places greater weight on unplanned hospitalization and ER use because these outcomes also capture the potentially avoidable use of hospitals' and ERs' resources. The Commission has underscored the importance of including both quality and resource use in measures of efficiency. A high rate of potentially avoidable adverse events indicates not only low quality but also inefficient use of hospital resources. By safely and appropriately preventing avoidable hospitalizations and use of the ER, home health care can efficiently reduce the use of hospital resources. The SQI score restricts the definition of adverse events to ER and hospital use for specific diagnoses that could have been prevented.

TABLE 4–3 **Agency level quality scores in the model**

	Agency					
	1	2	3	4	5	6
SQI score:						
Year 1	0.46	0.30	0.66	0.83	0.95	1.09
Year 2	0.60	0.61	0.69	0.86	0.87	1.16
Pooled data	0.50	0.56	0.66	0.85	0.92	1.13

Note: SQI (Standardized Quality Index).

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

Giving more weight to measures that include resource use is consistent with goals established by CMS and IOM. In the proposed demonstration of a P4P system in home health care, CMS has given additional weight to unplanned use of the hospital and use of the ER. In its report, IOM stressed the need for P4P to include measurements of resource use.

We discuss specific and additional issues in the development of composite quality scores at the end of this chapter.

Whether to measure quality for Medicare patients only

In our model, we measure quality only for Medicare patients cared for by Medicare-certified agencies. Choosing to measure only those patients for whom Medicare is the primary payer increases the homogeneity of the patients compared across agencies: Medicare patients tend to share certain characteristics such as age, full insurance coverage, and regular sources of care. Also, within home health care, patients must meet the same conditions of medical necessity and level of need: The rules of Medicare stipulate that home health patients must be homebound, require skilled medical services, and need temporary or intermittent care (rather than 24-hour or long-term care). Patients with non-Medicare sources of payment might not fit these criteria. The heterogeneity of private pay and Medicaid patients might make it more difficult to make fair comparisons of patients across agencies. In terms of the verification of data, patients outside of Medicare pose a special challenge because the Medicare program may not have a regular, auditable source of data for those patients.

Alternatively, a P4P system could include all of a provider's patients and not just those whose primary payer is Medicare. The Medicare program's conditions of participation maintain the same quality standards for all of a provider's patients. Some patients have both Medicare and Medicaid sources of payment; thus, the primary source of payment may change but the patient remains the same. Measures that are more inclusive allow for larger samples, which can result in more accurate quality measurement.

Illustration of a home health P4P model

For the model, we used the SQI score for the six agencies' therapy patients. This score summarizes 22 outcomes for patients who need physical therapy. Using primary diagnosis, which acts as a risk adjuster, we grouped similar patients together. Only Medicare patients are included.

On this scale, higher scores indicate that more patients achieved better outcomes more frequently. The scores ranged from –2 to +2. The average score was 0.84. The measurement periods are year 1 (from the second quarter of 2004 to the first quarter of 2005) and year 2 (from the second quarter of 2005 to the first quarter of 2006). Table 4-3 presents the average quality scores for the six agencies.

The third row in Table 4-3 displays each agency's score when we pooled data from year 1 and year 2. Pooling data across years is an effective tool to address the challenge of small sample sizes. Also, pooled data add stability to the scores because a two-year average changes less from year to year than a single-year average. As we continue to discuss the model in this chapter, we will measure these agencies by their score on the two years of pooled data.

TABLE 4–4 **The share of agencies in the reward group will depend on clinical group and statistical confidence level**

	Confidence level	
Clinical group	**95%**	**90%**
Therapy	32.0%	34.4%
Acute CVD	16.5	20.0
CHF or COPD	27.3	30.5
Diabetes	21.0	24.3
Pneumonia	15.3	18.9
Skin infection	16.7	20.5
Skin ulcer	14.6	18.3

Note: CVD (cerebrovascular disease), CHF (congestive heart failure), COPD (chronic obstructive pulmonary disease).

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

Setting thresholds for reward and penalty

P4P programs measure the quality of each provider and compare providers' quality scores with a threshold to determine whether they qualify for a reward for attaining high quality. Three components of the program can be set in advance: (1) the amount of the payment (necessary for budget-neutral systems), (2) the threshold that will trigger payment or penalty, and (3) the number of agencies that will receive a payment or a penalty.

For illustration, we have set both the funding and the threshold in advance. We call the threshold the national benchmark. Setting the quality target in advance may help some providers develop plans to improve quality, focus their efforts, and set milestones over the course of the measurement period to calibrate their performance. Alternative models that predetermine the proportion of agencies to reward or penalize (e.g., a system that rewards the top 10 percent or penalizes the worst 100 agencies) could penalize or reward average providers because some agencies that are statistically the same as the average could fall into the reward or penalty group. However, predetermining the size of the pool has the advantage of producing a stable, predictable pool of agencies to reward and penalize.

In comparing the agency's average quality score to the national benchmark, we use a statistically significant difference as the threshold: Thus, the threshold for a reward is to be statistically significantly above the national benchmark. The threshold for a penalty is to be statistically significantly below the benchmark and not show any year-to-year improvement. This system minimizes uncertainty by reducing the number of times it rewards a provider that is actually poor or mediocre or penalizes a provider that is actually mediocre or good.

The national average SQI score for therapy patients for the measurement year is 0.84 in the model. Whether a given agency is significantly better than average depends on three things: (1) the agency's score, (2) the size of the agency, and (3) the variation in outcomes among the agency's patients. High scores, larger samples, and more consistency increase the statistical certainty that an agency's score is greater than average; small samples and inconsistent outcomes among an agency's patients could lead to a score that is higher than average due to chance rather than to high quality of care.[7] Two sources of variation, measurement error and random variation in patients' response to care, could cause an agency's score to differ from the true quality of the agency.

The national average SQI score for therapy patients for the year before the measurement year is the benchmark of the system. This system would allow providers to know their quality improvement target; they would know what score they had to beat to gain a reward or how much they would need to improve to avoid a penalty. Thus, setting the benchmark with the previous year's average substantially reduces one of the greatest uncertainties providers in a P4P system face. Also, by using a national average the industry has already obtained, the program can be fairly certain that some providers will exceed the benchmark and some will fail to meet it. Alternatively, the trend in quality improvement that has emerged over the past several years of quality reporting in home health care—namely, about a 2 percent annual gain in functional outcomes—could be applied and the benchmark could be set 2 percent higher than the previous year's national score average so everyone would need to continue to improve at the current rate to maintain their current status; they would need to expend an additional effort to excel.

The reward group

When we apply the model to national data for patients in the therapy group, we find that we would place 34.4 percent of all agencies in the reward group (Table 4-4).

If we had started with a different clinical group, a different proportion of agencies would be eligible for a reward. Fewer agencies excel at care for the other six clinical groups. Also, if we applied a higher standard of certainty—for example, if we had used a 95 percent confidence interval—we would have a smaller proportion of agencies in the reward group.

Alternatively, P4P in home health care could use a model that is similar to the system CMS is considering for its home health P4P demonstration. This system will reward the top 10 percent of eligible agencies. This design has the advantage of ensuring that there will always be a group of agencies to reward. A system that sets a performance-based threshold runs the risk that very few or even no agencies will qualify for a reward. To be eligible, an agency must serve at least 25 patients. CMS's system measures all the patients at each agency. It does not restrict its measurement to patients in a single clinical group. As we noted previously, the CMS design scores each outcome separately; thus, an agency could receive a reward for its ability to improve patients' bathing but not receive a reward for improvement in walking.

A weakness of CMS's method of setting a threshold for reward is the potential to make statistical errors. Some agencies may score in the top 10 percent due to chance. Treating each agency's reported score as given—without accounting for the size of an agency's caseload or the standard deviation of scores within an agency's caseload—makes substantial distinctions among small agencies with widely variable scores and makes very little distinction among larger agencies with more stable scores that remain closer to the mean. The high level of variation in the scores of small agencies relative to the larger agencies indicates that their scores are likely to be the luck of the draw. They depend more on chance than on the underlying quality of the agency because the sample of patients is small. A threshold that ignores statistical significance would reward or penalize fewer large agencies with stable scores close to the mean and would reward or penalize more small agencies because of high variance in outcomes associated with small samples of patients. On the other hand, using a test of statistical significance implies that a large agency with a score close to the threshold may receive a reward while a smaller agency with a score well above the threshold would not receive a reward. One may wish to consider pairing a test of statistical significance with an absolute minimum difference from the threshold to limit the number of times very small but significant differences are rewarded.

The penalty group

For the purpose of the illustration, we set the threshold for penalty at a score statistically significantly lower than the national benchmark. The statistical method for determining this threshold is the same as the method we are using for the illustration to set the threshold for reward. In the illustration, we find that 28.9 percent of agencies fall into the penalty category. As in the case of the reward threshold, if different clinical groups were used, the proportion would be different, and, if we used a higher level of confidence, the penalty pool would be smaller.

Most P4P systems do not use penalties. There may be several reasons not to use them:

- Many P4P programs are voluntary; providers may be unlikely to volunteer for a program that could reduce their revenue.[8]
- P4P systems that are funded with generated savings or new money do not need a penalty pool to fund the rewards.
- Some suggest that the use of penalties will increase the amount of gaming that is likely to occur under a P4P system.

On the other hand, the possibility of a penalty is likely to motivate the providers in the middle and lower-middle portion of the quality spectrum to improve so that they may avoid losing revenue. A system without penalties might not provide enough motivation for some of the poorest performers to improve, because there would be no cost to them for nonparticipation.

The average group

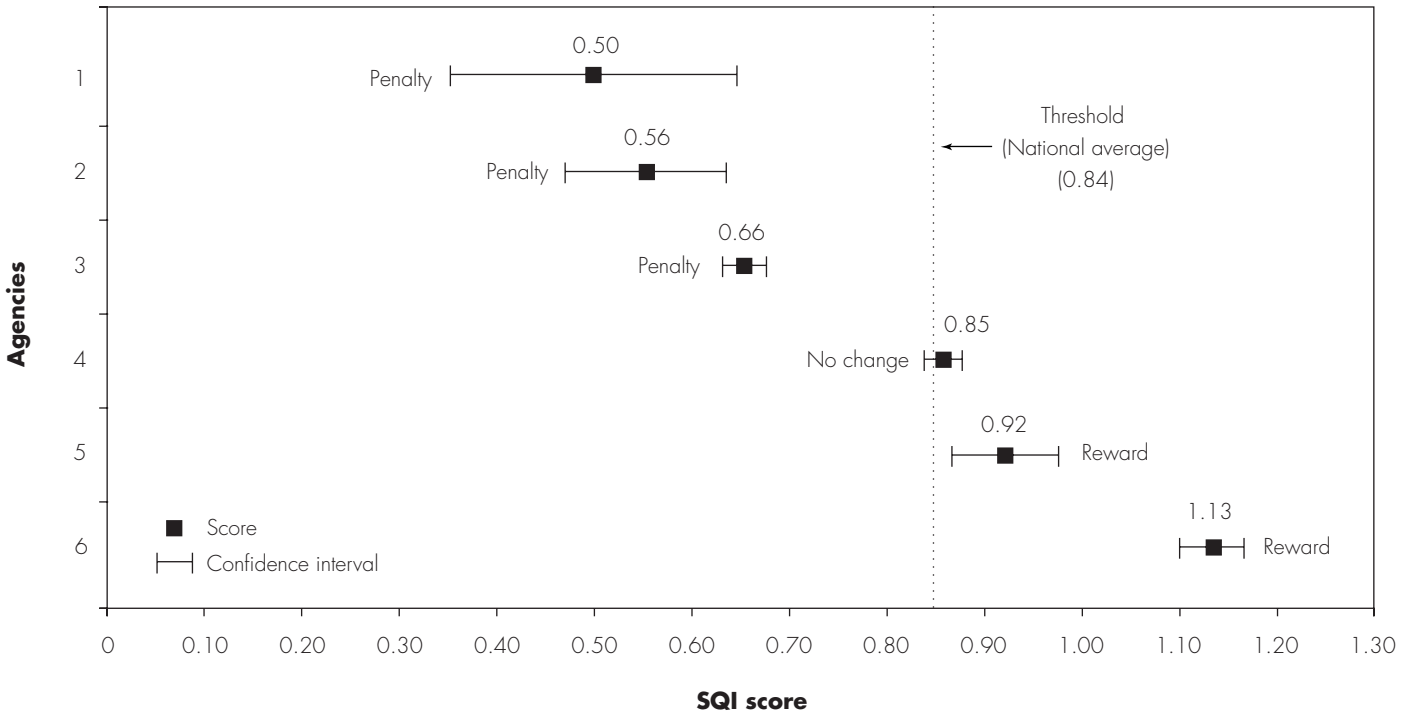
To illustrate how to apply thresholds, the model has a third group: agencies with neither reward nor penalty. They are neither statistically above nor below the benchmark. Not surprisingly, many agencies fit this category. In the model, they would receive a refund equal to the amount of payments withheld. However, these agencies may be eligible for a reward based on improvement, even though they do not attain high quality.

Illustration of a home health P4P model

For purposes of illustration, the threshold for reward is set at a level that is statistically higher than last year's national average; the threshold for penalty is statistically lower than the national average. The national average score was 0.84.

FIGURE 4–1

Comparing agencies to the threshold in the model



Note: SQI (Standardized Quality Index). The figure shows the agencies' pooled data score, which includes two years of data.

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

An agency whose confidence interval falls entirely below 0.84 is in the penalty group. If the confidence interval includes 0.84, the agency is in the no-change group. If the confidence interval is entirely above 0.84, the agency is in the reward group (Figure 4-1).

In the national data set, 34.4 percent of all agencies were eligible for a reward; a penalty was applied to 28.9 percent of agencies. In the proposed model, a third group of agencies (36.7 percent of the total) would be in neither the reward nor the penalty pool. Their scores are essentially the same as the average score; their quality is neither excellent nor poor.

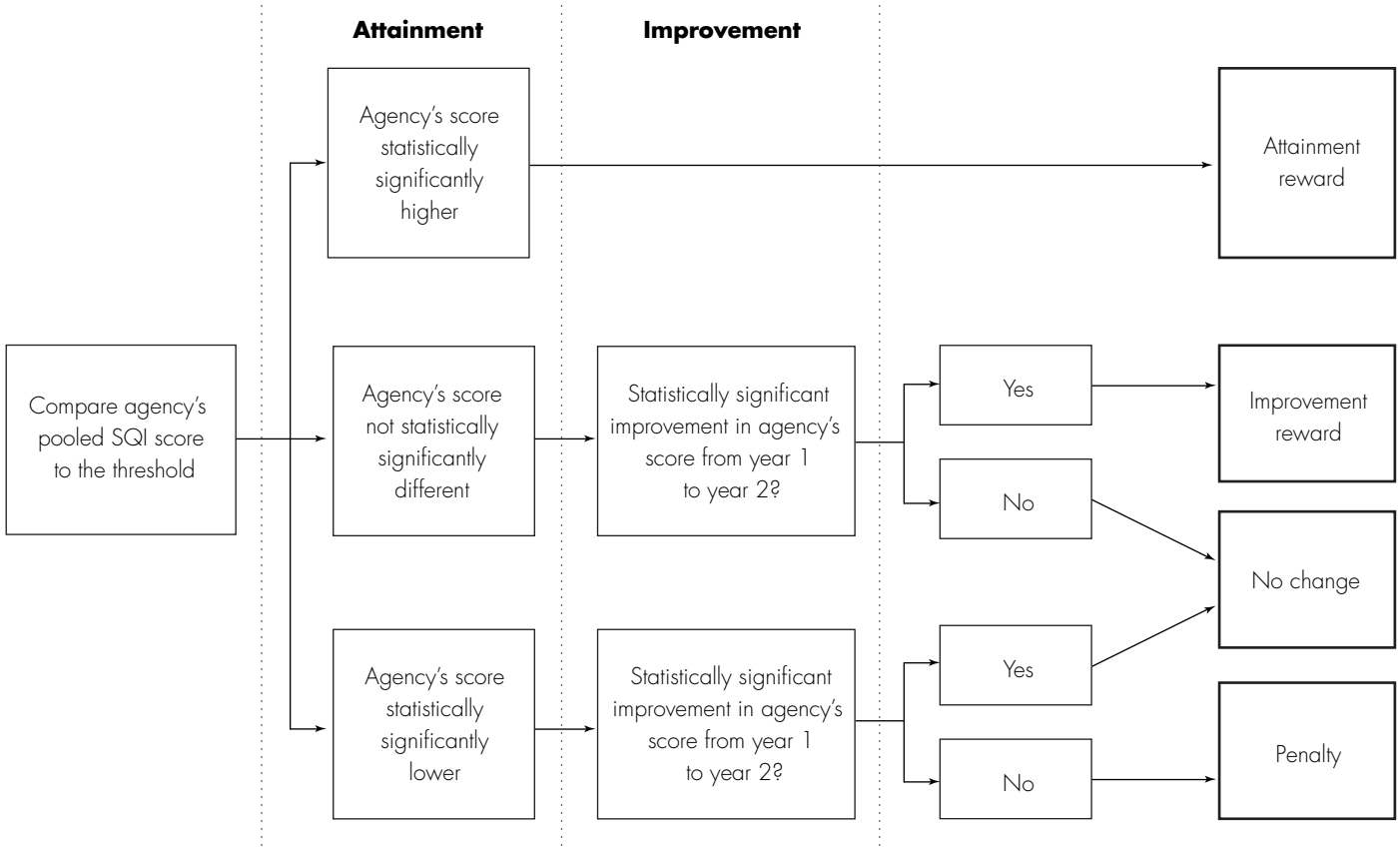
Balancing improvement and attainment

Next, the model considers improvement in agencies' performance over time, consistent with the Commission's principle that P4P should reward both attainment of high performance and improvement. In the model, agencies with average scores in the measurement year but with statistically significantly higher scores than they had in the previous year are eligible for an improvement reward. The award to this "most improved" group is half the size of the reward to the group that attained high scores. In the model, we also look again at the agencies in the penalty group. If they significantly improved over the previous year, they are lifted out of the penalty group and put into a group that receives neither reward nor penalty.

For the illustrative model, the second component of the reward system would acknowledge the improvement among agencies that did not attain a score high enough for an attainment reward (Figure 4-2). The Commission has stated as a principle that P4P should reward both attainment and improvement. If the improvement in an agency's score from the previous year to the current year is statistically significant, then that agency could be eligible for an improvement award. We use exclusive categories for attainment and improvement rewards. If an agency is eligible for an attainment award, it is not also eligible

FIGURE 4–2

Rewarding attainment and improvement in the model



Note: SQI (Standardized Quality Index). Agency's pooled SQI score includes two years of data.

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

for an improvement award. Thus, improvement rewards would go to agencies with average scores but that showed substantial progress toward the goal of excellence. In the model, the rewards for improvement would be one half the size of the rewards for attainment.

The illustrative model would reward agencies with average scores and any amount of statistically significant improvement. Agencies with scores that are statistically significantly below the benchmark would not be eligible for an improvement reward. Measuring the statistical significance of the difference in year 1 and year 2 scores would minimize the number of times we would give an improvement reward to small agencies with very unstable scores—agencies with scores that are likely to be higher or lower due to chance rather than to the influence of real quality improvement. Alternatively, there may be a minimum threshold for improvement such as a 10 percent difference between year 1 and year 2 so that small but statistically significant differences would not be rewarded.

The model also uses a measurement of improvement to soften the penalty for poor performance. If an agency's score were statistically significantly below the national benchmark score, but the agency showed significant improvement over its score the preceding year, then it would not receive a penalty. This system softens the penalty by allowing agencies who are truly getting better to avoid losing revenue. Thus, only the worst actors in the system would be penalized: They are both poor performers relative to the benchmark and are not getting any better relative to their own performance.

Illustration of a home health P4P model

In this step of the illustrative model, agency 3 avoids the penalty because its improvement from year 1 to year 2 was statistically significant. Agency 6 also had significant

improvement but has already qualified for an award based on its attainment, so in the model it cannot also receive the improvement reward. The other four agencies did not show significant improvement.

In national data, about 5 percent of all agencies that would have been in the penalty pool were lifted into the no-change pool because they showed significant improvement. Another group of 5 percent of agencies would be in the improvement group. They showed statistically significant improvement from year 1 to year 2, but in year 2 their score remained statistically similar to the average. These agencies would not qualify for an attainment reward but would qualify for an improvement reward. One could contemplate a further evolution of this scoring system in which agencies that attained a high level of performance with scores statistically significantly above the mean and also improved from year 1 to year 2 might be eligible for some additional bonus recognition as a breakthrough group.

Calculating rewards

The final step in the P4P system is distributing the rewards to providers. The Commission's principle that P4P should be budget neutral guides this step. The agencies in the penalty group will not have their 5 percent withhold returned to them. The 5 percent withhold is returned to the agencies in the no-change group. The agencies in the reward group receive an amount equal to the 5 percent withheld plus the reward amount. Because the model does not force the reward group and the penalty group to be the same size, and the pool was funded by a withhold of a predetermined size, the size of the rewards varies to fit the size and number of reward recipients. The amount returned or rewarded to an agency is proportional to the agency's Medicare payments. The size of the reward will also depend on the number and size of the agencies in the penalty group relative to the number and size of the agencies in the reward group.

Keeping the rewards proportional to Medicare's payments is consistent with our principle of realigning the payment system; that is, Medicare pays agencies in proportion to services rendered and so P4P rewards should distribute money under the same principle. However, the resources required to improve quality might not be proportional to revenue. If a minimum investment is required to achieve higher quality, then smaller agencies might need to commit a greater proportion of their resources than a larger agency. Establishing a minimum award amount may lead smaller agencies to believe the amount of the reward is a reasonable return on investment compared with the effort required to improve quality.

Illustration of a home health P4P model

In the model, we would be ready at this step to assign penalties and rewards to the six agencies (Table 4-5). The penalties against agencies 1 and 2 were withheld throughout the year. In the model, penalized agencies would not be required to pay the program any additional amount at the end of the year. Agencies 3 and 4 would receive a refund equal to the total amount withheld. Recall that agency 3 would have been penalized but it showed significant improvement and thus moved into the group that receives neither penalty nor reward. Agencies 5 and 6 would receive the reward payment calculated in Table 4-5 ($22,825 and $42,020, respectively) as well as a refund of the entire amount withheld ($20,800 and $38,200, respectively) for total year-end payments of $43,625 and $80,220, respectively.

Additional design features

The previous section summarizes the five important design features for a P4P program. In the process of building the illustrative model, we learned that we needed to address two additional features of the program—how to broaden the program to include the most agencies and how to improve the quality measures on which performance is rewarded over time.

Including providers with small numbers of patients

In the home health sector, like the other sectors of the Medicare program, a number of agencies will be too small to earn a reward or pay a penalty. In the illustrative model, because we consider sample size when we calculate statistical significance, many agencies will not be statistically distinguishable from the average. In alternative systems that compare scores with a threshold without considering statistical significance, there is generally a minimum sample size for inclusion and smaller providers are excluded from the system.

In the future, we could consider excluding agencies with a small number of Medicare patients from P4P. However, excluding small agencies introduces some perverse incentives that may run counter to the intent of the P4P system. An incentive that encourages low volume could create an access problem for beneficiaries. It could encourage medium-sized agencies to split or reorganize

TABLE 4–5 **Pay-for-performance reward and penalty amounts in the model**

	Agency					
	1	2	3	4	5	6
Total Medicare payments	$192,000	$755,000	$4,706,000	$2,106,000	$415,000	$764,000
Payment						
Penalty	–$9,600	–$37,700	$0	$0	$0	$0
Refund	$0	$0	$235,000	$105,000	$20,800	$38,200
Reward	$0	$0	$0	$0	$22,825	$42,020
Total	–$9,600	–$37,700	$235,000	$105,000	$43,625	$80,220

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

in ways that wastefully duplicate administration and overhead. It also removes the incentive for the system to develop new measures that could include smaller agencies.

Rather than exclude small agencies, a P4P system could address the issue of small agencies in at least two ways. One approach is to allow multiple small agencies that serve the same areas or contiguous areas to form voluntary quality associations. All the patients in the association would be pooled to count toward a single measurement. The association may generate a reward or a penalty. The agencies within the association could choose how best to distribute the results. This approach may encourage collaboration among agencies as well.

Another approach we found to be useful is to pool data for agencies across two consecutive years rather than use a single year of data for measurement. Pooled data yielded a substantially higher number of agencies with rewards and penalties. To be equitable, this pooling should be applied to all agencies and no one would have the opportunity to opt out of pooling. This approach has the additional advantage of resulting in more stable quality scores from year to year. It reduces the variation over time, the impact of small samples, and the potential impact of one-time events such as a change in management.

In the model, we had only the two most recent years of data, so when we measured improvement over time we used two scores, each based on only one year of data. A better alternative would be to use pooled data for the improvement score as well. The home health sector already has more than two years of data available, so pooling data over time would not necessarily postpone implementation of the program.

Improving the pay-for-performance measure set over time

In March 2005, the Commission suggested that additional measures be developed to complement those that have already been developed, collected, and used for quality measurement in home health care. The current set of measures focuses on the clinical effectiveness of care given to patients whose physical conditions are improving. Adding measures could broaden the patient population covered by the set, capture safety as an aspect of quality, capture a process of care directly under providers' control, reduce variation in practice, and provide incentives to improve information technology.

Apply process and safety measures. Process measures capture an aspect of care that is under providers' control: whether providers take very specific actions in the course of caring for their patients. Both the Commission and CMS have been considering adding process measures for home health care. The Commission convened a panel of researchers, quality measurement experts, and home health providers to identify best practices in fall prevention and wound care. Interest in these areas is high because falls and wounds are prevalent among home health care users. In addition, the practices are a part of the care for patients whose physical condition is not improving and for patients who are improving, and the practices are related

to patient safety (MedPAC 2006). CMS is working on developing other process measures.

The National Quality Forum also identified patient safety as an important dimension of quality—as outlined by IOM in its seminal study—and a priority area for quality measurement in home health care (IOM 2001).

As P4P begins to link reported quality levels with payment, the system should improve its ability to audit and verify the data. CMS has begun to develop these capacities within the Reporting Hospital Quality Data for Annual Payment Update program. Under this program, hospitals' quality data are audited to determine whether they are complete and whether they include a fair and sufficient sample of all their patients. Additional capacity to compare quality reports to other sources of administrative data or to audits of medical charts would further strengthen a P4P program. Adding process measures to the set of outcome measures for home health care would allow home health quality data to be verified through an audit of medical charts or through a comparison to information on the claim for payment.

Expand use of health information technology. The Commission recommended that P4P include measures of the functions supported by information technology (MedPAC 2005). Examples include a registry for patients with chronic conditions; a system that tracks test results; a system that can directly notify patients of laboratory test results; and a system that can aggregate, measure, and monitor patients by disease, medication, or other category. The functions of a telehealth system to remotely monitor patients' vital signs might be particularly relevant to home health care.

Furthermore, financial incentives for measuring and reporting care processes could encourage providers to improve their systems' capabilities to meet the new data requirements. When nurses, therapists, and other home health professionals are encouraged by best practices to assess, record, use, and share more information about patients' health status during an episode, wider use of information technology may result. These technologies include:

- ***Electronic medical records.*** The use of electronic medical records to store and provide information on a patient's past medical history, lab reports, and medications could greatly enhance the ability of health professionals to make informed decisions about care. In addition, electronic medical records allow an organization to measure its quality of care in real time rather than waiting for quarterly or annual measurements.
- ***Management tools.*** Patient registries, clinical reminder systems, and computerized patient assessments help providers manage a specific aspect of care.[9] If nurses used a computer program to help prompt and record patient assessments, it could reduce the burden of recording important clinical information, suggest appropriate tests, and immediately identify patients who need special interventions to address their needs.
- ***Patient communications.*** Devices used in patients' homes to monitor their health can make it easier for patients to monitor their condition, communicate with caregivers, and identify the need for a medical intervention.

Patient experience measures. Many agencies already collect patient satisfaction information. A basic patient experience questionnaire might not be radically different from activities many agencies already conduct. If the program wished to phase in patient experience measurement, it could begin with a pay-for-reporting step in which all agencies would have the incentive to develop or hire the capacity to survey their patients.

A standardized tool that could be audited and administered with some independence from the agency staff being evaluated would be necessary to compare patient experience measures among agencies. Potential patient experience measures include:

- How often did nurses listen carefully to you?
- How often did nurses explain things in a way that you could understand?
- How often was your pain well controlled?
- Did you get information about symptoms to watch for after you were discharged?

As this partial list suggests, patient experience measures can begin to capture concepts such as the adequacy of planning for patients' transitions from professional home health care to living in the community or concepts such as the patient-centeredness of care (whether patients feel adequately informed to actively participate in their care).

Circumstances of the home health sector

Though the P4P framework discussed in this report would realign some funds for incentives to reward quality, most Medicare payments for home health care would still be administered under the provisions of the current prospective payment system (PPS). MedPAC and others have cited issues with the PPS, and some of these issues could diminish the impact of a P4P incentive (MedPAC 2006, GAO 2000). Adding a quality incentive to a payment system that does not accurately pay providers for the costs of different patients could create perverse incentives for providers—or overpower the impact of the quality incentive. Many factors suggest that the current system overpays providers and pays inaccurately for some patients.

Concerns about payment accuracy underscore the need to use P4P in tandem with other efforts to reform the home health payment system. A quality incentive will redirect funds toward a defined outcome that is valuable to beneficiaries and improves the incentives under PPS. However, maintaining incentives for efficiency under the core PPS is critical. Improving quality without maintaining incentives for efficiency could cause a conflict between efforts to improve quality and efforts to address Medicare's long-term sustainability challenge. Continuing efforts to improve the accuracy of payments under the PPS will ensure that providers have appropriate incentives to provide quality care.

The aggregate average financial performance of the home health industry under PPS has been remarkable (MedPAC 2006). Since the advent of the PPS, most agencies have held per episode cost inflation to about 1 percent per year, and margins have exceeded 10 percent despite a one-time reduction in the base rate and numerous reductions to the update. The consistent pattern of high margins suggests that the base payment in the home health PPS may not accurately reflect the costs of efficient providers, potentially dimming the impact of a reward or penalty for quality. For agencies with significant margins, such as the 50 percent of agencies with margins greater than 16.8 percent in 2007, the impact of a 5 percent reward or penalty may be too modest to encourage quality improvement.

Shortcomings in the case-mix measurement may provide incentives for HHAs to favor patients with higher case-mix scores. Prior analysis has found a small but statistically significant relationship between an agency's case mix and its margins (MedPAC 2005). Medicare's system for classifying patient resource needs, the home health resource groups (HHRGs), may also inappropriately group patients within a single case-mix group though they have very different resource needs. MedPAC found a large variation in the minutes of service per episode provided to patients in the same HHRG (MedPAC 2006). The case-mix weights for home health care have never been updated, and as a result it is unlikely the current case mix accurately reflects the resource intensity of different patients.

Differences in financial performance among providers are to be expected in any PPS, as providers vary in their efficiency. However, if some of this variation in margins is due to the issues highlighted above, then the variation reflects shortcomings in the PPS. This variation may affect a quality incentive because providers are likely to assess the value of any incentive relative to their margins. For example, the top quarter of HHAs, which have margins that exceed 27 percent, might not consider a 5 percent incentive compelling. Medicare should not expect the margins of providers to necessarily be concentrated, but failing to address inaccuracies in the payment system that can lead to excessive variation may diminish the impact of a quality incentive.

Additional technical information on home health pay for performance

In this section, we discuss some limitations of the risk adjustment currently available for home health outcome measures, the composite measure we developed to summarize quality at the agency level, and adjusting for socioeconomic status.

Adequacy of risk adjustment for home health measures

CMS developed risk adjustment for the OBQIs to take into account patient health and other characteristics that may affect their outcomes. For example, improving patients' pain from cancer is more difficult than improving pain in patients with congestive heart failure because of the extreme pain associated with many cancers. Early studies found that risk adjustment was accounting for the impact of patients' primary diagnosis on pain and giving "credit" for the difficulty of cancer patients' pain management. In essence, taking these patient characteristics into account should level the playing field among agencies with different patient populations.

However, when we applied the risk-adjustment methodology that was calibrated in 2001 to the most recent data available from 2005, we found that it did not adequately account for differences in patient mix at the agency level (Shaughnessy et al. 2002). Some of the limitations of CMS's risk model might be explained by the fact that it has not been recalibrated since the measures were implemented more than five years ago. In the calibration year, the expected values and the actual values were almost the same. As time passed, the gap between the model's expected values and the actual values widened. For example, by 2005, the predicted rate of success in improvement in ability to dress the upper body was 60 percent, and the actual national rate was 67 percent. If the changes that led to the gaps in the model's performance have not been consistent among patient types, that would explain the model's limitations in predicting current outcomes by patient type.

Our two tests of the risk-adjustment system applied to the most recent available data suggest that the risk adjustment does distinguish between patients with very low likelihood of good outcomes and those with very high likelihood of good outcomes. However, the system is not as capable of making finer distinctions. The risk adjustment correctly identifies the general patterns in outcomes, but it is not very precise.

In one test, we divided the patients into deciles (10 groups of equal size). The groupings were based on CMS's risk-adjustment model's prediction of the relative likelihood of their success at the outcome we were measuring. In each test, the model predicts the broad pattern in the relative rate of success for patients: Those in deciles with the lower predicted rates of success do achieve lower rates of success than those in higher deciles. However, the risk-adjustment model is imprecise; there is often a wide gap between the predicted rate and the actual rate.

In another test, we found that the risk-adjustment model did not precisely account for differences in outcomes that were related to patient characteristics. We considered patient characteristics such as primary diagnosis, comorbidities, informal caregiver availability, and functional limitation. We chose these characteristics because previous research indicated that they are likely to influence outcomes (Shaughnessy et al. 2002).

We found statistically significant differences among the outcomes for different patient types after we applied the risk-adjustment model. In other words, though we had tried to account for the effects of each of the patient characteristics in our expectations, we still found that patients of certain types had much better outcomes than patients of other types. The results of this second test reinforced the evidence from our first test: The CMS risk-adjustment model seems to have some limitations in its ability to level the playing field among different types of patients. Even with risk-adjusted data, many differences will exist between the outcomes of patients of different types. This will reduce the validity of the quality score, will give an advantage to agencies with certain mixes of patients, and could lead to access problems for patients of certain types.

A composite home health quality measure to combine measures of quality and address shortcomings in risk adjustment

A composite can bring several measures together to create a picture of quality that is more complete than a single measure can be. Any single measure of quality excludes some providers, some patients, or some trait of quality. We studied quality composites from scorecards for hospitals from states and private plans and worked with technical experts to develop potential criteria for good composite measures. The composite measure should:

- apply to most providers, most patients, and most quality traits;
- account for differences in patient characteristics;
- reflect the relative importance of each measure in the composite;
- be easy to describe and understand; and
- acknowledge the extent of uncertainty and identify where it exists.

Both the selection of measures to include in the composite and the construction of the composite determine whether the composite meets the criteria.

We contracted with a quality benchmarking organization to help us construct a composite measure for HHAs. They applied expertise in clinical logic, statistics, and measure design to the national data set of all OASIS patient assessments to develop a composite quality measure: the SQI. The SQI is risk adjusted by clinical stratification instead of by CMS's regression-based system. This allows us to identify a relatively homogeneous set of patients at

each agency and compare each agency's score for those patients rather than rely on risk adjustment to account for all the differences among all of each agency's patients.

Clinical stratification groups patients with similar diagnoses and treatment plans. This allows the measurement system to compare the outcomes for similar patients at different agencies. It also establishes a clear link between patient groups and outcome for the agency. If an agency wishes to target a particular outcome, the measurement system has already identified the patients and treatment plans that need to be addressed. However, clinical stratification is generally regarded as incomplete risk adjustment because of the variables it does not address. In the long run, CMS may wish to explore a hybrid model that groups patients into clinical classifications and also applies regression-based risk adjustment within groups to account for additional sources of variation.

The SQI measure relies on the OASIS patient assessments performed by home health nurses and therapists at admission, at some intervening events, and at discharge to determine the outcomes of patients' home health care: whether patients' functional levels improved or stabilized and whether patients experienced any adverse events. The components of the measure are detailed in Table 4-6.

The SQI set incorporates the seven publicly reported functional measures from the Home Health Compare public data report, adds more functional outcome measures, and adds the four potentially avoidable adverse events listed in Table 4-6. These are gross measures not of all hospital and ER use but of that specifically due to four events the agency is thought to be able to manage.

We tested the correlations among the components of each measure. Using the statistical measure Cronbach's alpha, we determined that relationships among the constituent measures of each measure were acceptable. This statistical measure indicates the extent to which a set of test items can be treated as measuring a single construct. In this context, we are measuring whether we should use a set of functional outcomes and adverse events together to measure the quality of an HHA. We compared the SQI with an alternative measure that was limited to the public data report measures. We found an alpha of 0.71 for the measures in the SQI and an alpha of 0.60 for the measures in the simpler alternative. The alpha score for the SQI exceeds the rule-of-thumb standard for reliability of 0.70 (Streiner and Norman 1989). The lower score for the simpler alternative suggests that adding the additional components to the SQI is an improvement.

TABLE 4–6 **Components of MedPAC's quality score for home health pay for performance**

Functional outcome measures	Potentially avoidable event measures
• Getting out of bed • Walking • Bathing • Using the toilet • Urinary incontinence • Bowel incontinence • Upper body dressing • Lower body dressing • Shortness of breath • Caregiver managing medical equipment • Managing oral medications • Managing inhaled medications • Managing injectable medications • Managing medical equipment • Ulcer, stasis • Ulcer, pressure • Surgical wound • Pain • Confusion • Anxiety	Unplanned hospitalizations or uses of the ER caused by: • Diabetes out of control • Injury caused by a fall at home • Wound infection or deterioration • Improper medication administration

Note: ER (emergency room).

The steps to calculate an agency's SQI score are fairly simple. The system starts at the patient level. For each patient, all the functional outcomes are scored 2 points for improvement, 1 point for stabilization, and –1 point for decline. The scores for all the functional outcomes are summed and a point is subtracted for each incidence of a potentially avoidable unplanned hospitalization or ER use. The resulting total is divided by 20 to obtain an average. Finally, the scores for all of the patients in an agency are averaged. In our data, agencies' SQI scores range from –4 to +2.

Some patients who qualify for the home health benefit have limited potential for improvement. In the illustrative measure, points are available for stabilizing patients whose illness or functional level otherwise could have declined. The measure also includes a penalty for potentially avoidable hospitalizations and use of the ER, which has the effect of rewarding agencies who manage patients with

TABLE 4–7

Nearly all home health agencies treat patients in selected clinical groups

Clinical group	Agencies with more than: 2 patients in clinical group	25 patients in clinical group
Acute CVD	6,360	1,040
CHF or COPD	7,710	4,520
Diabetes	7,240	2,610
Pneumonia	5,980	1,070
Skin infection	6,870	1,520
Skin ulcer	6,510	1,450
Therapy	7,530	4,940

Note: CVD (cerebrovascular disease), CHF (congestive heart failure), COPD (chronic obstructive pulmonary disease). Between 2003 and 2005, there were about 8,000 agencies in Medicare.

Source: Outcome Concept Systems analysis of 2003–2005 cost report and Outcome and Assessment Information Set data.

declining health safely in their homes while preventing unnecessary hospitalizations and trips to the ER.

The SQI score incorporates steep penalties for unplanned hospitalization and ER use to reflect the importance of these measures as adverse events—and thus indicative not only of low quality but also of actual harm to beneficiaries—and measures of the efficiency of home care. One of the most important contributions home health care spending can make to the efficient resource use of the Medicare program is to safely and appropriately prevent avoidable hospitalizations and use of the ER. The Commission has underscored the importance of including both quality and resource use in measures of efficiency. For these reasons, the score is designed to give additional weight to adverse events. The SQI score restricts the definition of adverse events to ER or hospital use for four reasons: diabetes out of control, injury caused by fall, wound infection, and improper medication use. These four reasons describe events that were potentially preventable.

We calculate an agency's SQI for patients within a clinical group. Because the evidence reviewed in the previous section demonstrates that CMS's risk-adjustment model does not sufficiently account for differences in patients' outcomes based on their primary diagnosis, we chose to stratify patients into groups based on their primary diagnosis using the clinical classification system. We applied factor analysis to our large database to identify seven categories that included most patients and that put them in clinically related groups (patients who would receive similar treatments during the course of their home health care). The clinical classifications are listed in Table 4-7. Most agencies treat patients in these common clinical groups.

This measure is not as simple as an "off-the-shelf" solution, but it better meets the criteria for good measures that we have developed and discussed. The SQI is applicable to most providers, most patients, and most quality traits. The stratification into clinical groups accounts for differences in patient characteristics. The scoring method reflects the relative importance of improvement, stabilization, and adverse events for each measure in the composite. In our P4P model, we show how the SQI can be used to describe the extent of uncertainty and identify where it exists. Finally, the measure uses data that are part of the currently collected home health data.

Basing patient groups on primary diagnoses makes a clear link between patient groups and outcome for the agency: The measurement system identifies patients with similar treatment plans that need to be addressed. Focusing the P4P program on one group of patients or on several groups of patients provides guidance to agencies on how to focus their quality improvement efforts and might decrease the burden compared with a program that started with all of an agency's patients. However, relying solely on primary diagnosis for risk adjustment is generally regarded as incomplete risk adjustment because of the variables it does not address. In the long run, CMS may wish to explore a hybrid model that groups patients by primary diagnosis and also applies regression-based risk adjustment within groups to account for additional sources of variation.

Accounting for differences in socioeconomic status

In a program as comprehensive as Medicare, there may be wide differences in the socioeconomic status (SES) of patients in addition to differences in the clinical characteristics we have discussed thus far. Some suggest that socioeconomic differences among patients may lead to differences in the quality of care measured at the provider level for reasons beyond agencies' control. Patients in a lower socioeconomic group may lack access to competent informal care, may have fewer tools to make informed decisions, or may have a poorer quality diet than those of higher SES. However, deciding whether and how to adjust for socioeconomic differences is difficult.

Choosing whose socioeconomic traits, which traits, and what scales to use to measure SES can be challenging. In home health care, the characteristics of the patient's family might be as important as, or even more important than, those of the patient. This raises the question: Whose status should be measured—the patient, the immediate family, or the extended family?

There is some room for doubt about the relationship between SES and health outcomes. A recent study on breast cancer mortality found higher rates of mortality among women in higher socioeconomic groups than in lower ones (Strand et al. 2007). Another study found that much of the relationship between SES and health is a function of known health factors, such as obesity and smoking, which are measured directly and accounted for in the clinical risk adjustment (Kuper et al. 2007). SES may relate to different measures in different ways: It may have little impact on a process measure such as giving hospitalized patients an aspirin but it may have a larger impact on whether patients will purchase and consistently use medications to manage blood pressure after they return home.

Finally, adjusting for SES has the effect of setting lower expectations for the providers who are in a position to have the greatest impact on vulnerable populations. For example, if a Medicare P4P program were to use an SES adjustment that incorporated race, it could have the effect of setting a lower expectation for quality of care delivered to blacks than for whites, Hispanics, or other racial groups. Some may view lower standards for the care of vulnerable populations to be one of health care's critical problems; the impacts of disparities in health care have been widely studied. A P4P system that expects good care for all patients regardless of race, income, or education could be one policy tool to address the issue of disparities in health care.

Despite the difficulties associated with measuring SES and establishing its relationship with health outcomes, some contend that P4P should be used to address disparities in health care (Rosenthal and Dudley 2007). One approach for the future is to develop direct measures of health care disparity that can be attributed to providers and patients and reward providers for addressing it. Another approach to consider—using currently available measures—is to offer greater incremental payments to providers who achieve high quality for underserved populations. This would have the effect of increasing the incentive to better serve vulnerable beneficiaries as well as providing some adjustment to acknowledge that achieving high quality for underserved populations could require a greater effort than achieving these goals among other populations.

An alternative to SES-based adjustments to risk scores would allow providers to identify noncompliant patients and exclude them from their data. The United Kingdom uses this system in its nationwide physician quality incentive program (Doran et al. 2006). A comprehensive study of this design option found that most physicians exempted few of their patients. There was some evidence of abuse at the extreme, and they found a moderate correlation between the number of patients exempted and the quality score achieved by the physician. However, the opportunity that exception reporting presents to manipulate quality scores could be counterbalanced by publicly reporting the providers' noncompliance rates, auditing providers with exceptionally high rates, or requiring providers with a noncompliance rate above a certain threshold to develop and implement a plan to increase compliance. ■

Endnotes

1 The Institute of Medicine and CMS have also considered funding P4P through savings generated by quality improvements.

2 One survey of private-sector efforts found that purchasers report needing to provide incentives of 5 percent to 20 percent for physicians and 1 percent to 4 percent for hospitals (MedVantage 2004). Yet, it is difficult to know what portion of overall payment these percentages represent. Because Medicare payment is often a higher percentage of any one provider's total revenue than a single private payer, a smaller percentage of Medicare's payment may encourage change. In CMS's Premier hospital demonstration, preliminary results show improvement in all conditions in the first four quarters in anticipation of financial rewards of 1 percent or 2 percent for those in the upper rankings (Remus 2005).

3 Numerous studies suggest that patients frequently do not receive evidence-based care and often experience illness or injury as a result of contact with the medical system (Jencks et al. 2003, McGlynn et al. 2003, IOM 2001).

4 Both the study by the Premier group and a later study by a group of researchers outside of the system found greater improvement among hospitals within the demonstration than in hospitals outside the demonstration (Lindenauer et al. 2007). The Premier study was very positive about the implications of the results of the demonstration for P4P. The outside researchers concluded that the quality differences were small compared to the costs of operating the quality incentive program and suggested that the demonstration has negative implications about the cost effectiveness of P4P on a larger scale.

5 Based on MedPAC analysis of freestanding agencies' cost reports, in 2005, 5 percent of agencies received less than $125,000 and 5 percent of agencies received more than $6.5 million. The smallest agency in terms of Medicare revenue received $2,500 and the largest received $18.4 million.

6 Research that supports the reliability of OASIS items was conducted on the research and development sample of OASIS data. Later tests on OASIS from the field indicate lower levels of reliability for some items (Kinatukara et al. 2005).

7 Conceptually, we are treating each agency's case load for the measurement year as if it were a sample of patients drawn from the population of all patients at all agencies and measuring the sample mean, sample size, and standard deviation of scores within the sample. We are testing whether it is likely that the sample's average score is higher or lower than the population's average score due to chance or whether the sample is really different from the population; theoretically, it would be different because the quality of the agency is truly good or truly bad. We chose to apply a two-stage, one-tailed test of significance at a 90 percent level of confidence. We determine whether each score that is higher than the benchmark is significantly higher in stage 1, and then we determine whether each score that is lower than the benchmark is significantly lower in stage 2. For each of these two tests, we apply a 95 percent confidence coefficient.

8 In the case of CMS's proposed home health P4P demonstration, for example, the designers thought a penalty was not consistent with voluntary participation. We note, however, that CMS's hospital P4P demonstration was also voluntary and it did incorporate the possibility of a penalty.

9 These management tools are often embedded in an electronic medical record; however, they are also available on their own.

References

Doran, T., C. Fullwood, H. Gravelle, et al. 2006. Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine* 355, no. 4: 375–384.

Dudley, R. A., A. Frolich, D. L. Robinowitz, et al. 2004. Strategies to support quality-based purchasing: A review of the evidence. *Technical Review* 10. Prepared by the Stanford–University of California San Francisco Evidence-based Practice Center under contract no. 290–02–0017. AHRQ publication no. 04–0057. Rockville, MD: AHRQ. July.

Government Accountability Office. 2000. *Medicare home health care: Prospective payment system will need refinement as data become available*. GAO/HEHS–00–9. Washington, DC: GAO.

Institute of Medicine. 2006. *Rewarding provider performance: Aligning incentives in medicine*. Washington, DC: National Academies Press.

Institute of Medicine. 2001. *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Integrated Healthcare Association. 2006. *Advancing quality through collaboration: The California pay for performance program*. Oakland, CA: Integrated Healthcare Association. February. http://www.iha.org.

Jencks, S. F., E. D. Huff, and T. Cuerdon. 2003. Change in the quality of care delivered to Medicare beneficiaries, 1998–1999 to 2000–2001. *Journal of the American Medical Association* 289, no. 3 (January 15): 40–45.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47, no. 2 (March): 263–292.

Kinatukara, S., R. Rosati, and L. Huang. 2005. Assessment of OASIS reliability and validity using several methodological approaches. *Home Health Care Services Quarterly* 24, no. 3: 23–38.

Kuper, H., H. O. Adami, T. Theorell, et al. 2007. The socioeconomic gradient in the incidence of stroke: A prospective study in middle-aged women in Sweden. *Stroke* 38, no. 1 (January): 27–33.

Lindenauer, P. K., D. Remus, S. Roman, et al. 2007. Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine* 356, no. 5 (February 1): 486–496.

McGlynn, E. A., S. Asch, J. Adams, et al. 2003. The quality of health care delivered to adults in the United States. *New England Journal of Medicine* 348, no. 26 (June 26): 2635–2745.

Medicare Payment Advisory Commission. 2006. *Report to the Congress: Medicare payment policy*. Washington, DC: MedPAC.

Medicare Payment Advisory Commission. 2005. *Report to the Congress: Medicare payment policy*. Washington, DC: MedPAC.

Medicare Payment Advisory Commission. 2004. *Report to the Congress: Medicare payment policy*. Washington, DC: MedPAC.

Medicare Payment Advisory Commission. 2003. *Report to the Congress: Medicare payment policy*. Washington, DC: MedPAC.

MedVantage, Inc. 2004. *Pay for performance programs for providers increase dramatically in 2004*. Press release. San Francisco, CA: MedVantage. December 15.

Premier, Inc. 2006. Centers for Medicare and Medicaid Services (CMS)/Premier hospital quality incentive demonstration project: Findings from year one. Charlotte, NC: Premier, Inc. http://www.premierinc.com.

Remus, D. 2005. Presentation at CMS/National Quality Forum implementing NQF-endorsed consensus standards meeting. Implementing a hospital-based pay-for-performance model: Challenges and opportunities. May 9.

Rosenthal, M. B., and R. A. Dudley. 2007. Pay-for-performance: Will the latest trend improve care? *Journal of the American Medical Association* 297, no. 7 (February 21): 740–744.

Rosenthal, M. B., Frank, R., Li, Z., et al. 2005. Early experience with pay for performance. *Journal of the American Medical Association* 294, no. 14: 1788–1793.

Shaughnessy, P. W., D. F. Hittle, K. S. Crisler, et al. 2002. *OASIS and outcome-based quality improvement in home health care: Research and demonstration findings, policy implications, and considerations for future change.* Vol. 2, Research and technical overview. Denver, CO: Center for Health Services Research, University of Colorado Health Sciences Center.

Strand, B. H., A. Kunst, M. Huisman, et al. 2007. The reversed social gradient: Higher breast cancer mortality in the higher educated compared to lower educated. *European Journal of Cancer* 43, no. 7 (May): 1200–1207.

Streiner, D. L., and G. L. Norman. 1989. *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press.

Town, R., D. Wholey, J. Kralewski, et al. 2004. Assessing the influence of incentives on physicians and medical groups. *Medical Care Research and Review* 61, no. 3 supplement.